Audio Engineering Society

Convention Paper 6407

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

The Intelligent Artificial Vocal Vibrato Effecter using Pitch Detection and Delay-Line

Chulwoong Jeon[1], Peter F. Driessen[2]

[1, 2]Dept. of Electrical and Computer Engineering, University of Victoria, P.O. Box 3055 STN CSC, Victoria, B.C. V8W 3P6, CANADA
[1]cwjeon@ece.uvic.ca, [2]peter@ece.uvic.ca

ABSTRACT

Karaoke is one of the largest commercial industries using audio & video in Asia today. One of the most popular features incorporated into the vocals produced by Karaoke singers has been Echo/Reverb. In addition to Echo/Reverb, if vibrato is added to vocal signals, then the vocal vibrato produced has the potential of making the singer feel more comfortable, confident and professional with regards to their singing. In this paper, we present a real-time 'Vocal Vibrato Effecter' running under a Windows PC, which automatically adds a vibrato effect to the vocal input. The proposed technique exploits the vocal energy level and the temporal consistency of the pitch variation. The key novelty in this work is the combination of pitch detector and pitch shifter. This effecter can be applied to consumer/commercial Karaoke systems to enhance a vocal signal.

1. INTRODUCTION

Karaoke, "sing along" with video, is one of the most popular forms of light entertainment in Asia. Karaoke equipment today consists of five functional parts [7].

1. Sound Source
2. Visual Source
3. Microphone
4. Content Distribution
5. Accessories (Vocal Echo, Automatic Scoring system, etc.)

All five elements have evolved along with the technological improvements made throughout the years. For example, the 'Sound Source' has been switched from analog tape to LD/CD, the 'Microphone' has been changed from wired to wireless; however, there have not been many improvements on the vocal effects component of the system. Due to a destitution of improvements on the side of vocal effects, we are inclined to review and survey new effects and improvements that can be added to this element.

When we listen to performances of professional singers, we often notice that the singer is frequently using 'Vibrato' to separate his voice from the background music and give richness to a tone being sung [1]. While monitoring vocal vibrato, we observed that a professional singer usually triggered the vibrato after holding an initial pitch for a certain time period (i.e. 200ms, 700ms, etc). Based on this observation, we developed the 'Vocal Vibrato Effecter' that has real-time processing capability.

2. RELATED SYSTEMS IN THE LITERATURE

The most similar work that could be found with regards to this system was in "A System for Hybridizing Vocal Performance"[5] proposed by Lau. In order to generate output sound, Lau's model requires two special elements that are 'Target Vocal Signal' and 'Dynamic Time Wrapping (DTW)' (Figure 1).

In Lau's system configuration, a singer's singing is synchronized to the 'Target Vocal Signal' at the 'Dynamic Time Wrapping (DTW)' in Figure 1. Also, a singer's singing and the 'Target Vocal Signal' are analyzed in the form of pitch and amplitude simultaneously.

Based on the pitch/amplitude information of a singer's singing/'Target Vocal Signal' and the timing information generated by DTW, 'Generation of Modification Parameters' creates transformation information. Finally, a singer's singing is transformed along with the transformation information.

Karaoke is a cost sensitive market; therefore, those two special elements in Lau's work cause limitations in terms of being practically applied as a real-time processing application for Karaoke systems.

First, as we reviewed, it requires the 'Target Vocal Signal' and it cannot produce the synthesized output without the 'Target Vocal Signal'. In addition, the 'Target Vocal Signal' has to be provided as a waveform segments to the 'Generation of Modification Parameters'. This method unfortunately creates a huge amount of data which requires equally huge storage space, which basically means greater costs in implementation.

Using the observation that we mentioned earlier, we are able to overcome the limitations existing in Lau's work. These new methods will be discussed in the remainder of this paper.

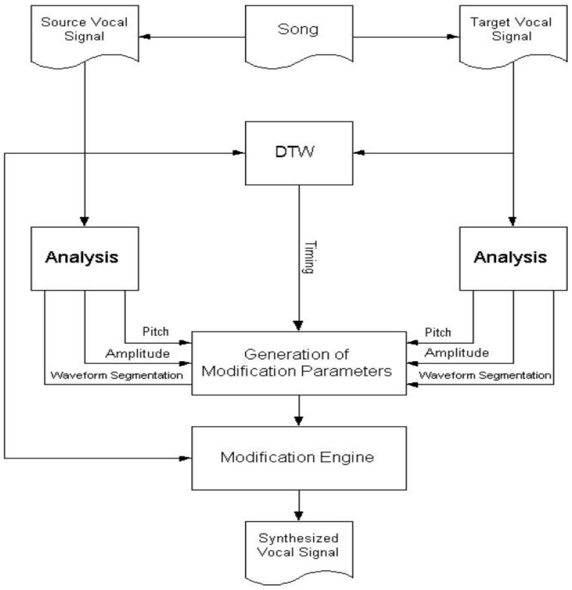


Figure 1 The System Configuration of Lau's [5]

3. SYSTEM CONFIGURATION

Based on the observation in Figure 1, we designed the following system (Figure 2).

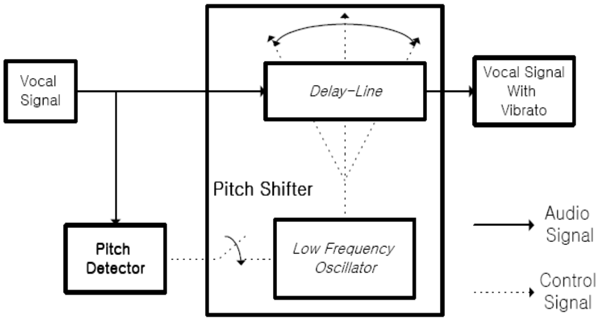


Figure 2 Intelligent Artificial Vocal Vibrato Effecter System Configurations

As Figure 2 shows, this effecter is composed of two parts. One is the Pitch Detector and the other is the Pitch Shifter with Low Frequency Oscillator (LFO).

The Pitch Detector has an internal buffer to hold previous samples with current ones. At specified times, the Pitch Detector evaluates the samples that have been held in the internal buffer in order to detect the fundamental frequency which are present in the pitch of singing and to measure the power. If the measured pitch stays within ± 1 semitone for a defined period of time and the measured power is higher than the threshold level (i.e. –40 dB), then the Pitch Detector turns on the LFO in the Pitch Shifter and the LFO begins to modulate the Delay Line (Vibrato starts). Meanwhile, if the measured pitch exceeds ± 1 semitone or the measured power is smaller than the threshold level, the LFO is turned off and stops the modulation of the Delay Line (Vibrato stops).

Three parameters, Latency Time, Vibrato Depth and Vibrato Speed, control this system. Latency Time is the window of time the Pitch Detector monitors the input pitch variation. If the input pitch variation stays within ± 1 semitone in this window, the Pitch Detector turns on the LFO in the Pitch Shifter. Vibrato Depth controls the pitch variation of vibrato output and Vibrato Speed controls the speed of vibrato output.

By combining those three parameters, our system can generate a Vocal Vibrato effect intelligently without supplying any external signal or data.

4. PITCH SHIFTER

If we vary time delay periodically this will represent a periodic pitch variation, which is called 'Vibrato' [2]. Using this phenomenon, the Pitch Shifter is designed using the Delay Line with the LFO [6]. Typically, the vocal vibrato occurs at a rate of 4 to 7 Hz and with a fundamental frequency modulation extent of ± 1 semitone [1]. If the Delay Line's length is a few milliseconds, then it is modulated by the LFO with 4 to 7 Hz speed, which will generate an effect similar to vocal vibrato.

5. PITCH DETECTOR

The Pitch Detector plays a very important role in our system (Figure 2) because it makes the decision of when the vibrato for the vocal signal should be turned on or off. There are several techniques to measure the pitch and we evaluated two methods: Harmonic Product Spectrum (HPS) which measures the pitch in the frequency domain and Autocorrelation which measures the pitch in the time domain in order to pick up the best one for the system.

5.1. Harmonic Product Spectrum (HPS)

The Harmonic Product Spectrum (HPS)[4] is looking for the maximum coincidence about harmonics using equation (1) for each spectral measurement $X(\omega)$

$$Y(\omega) = \prod_{r=1}^{R} \left| X(\omega r) \right| \tag{1}$$

$$\hat{Y} = \max_{\omega_i} \left\{ Y(\omega_i) \right\} \tag{2}$$

where $X(\omega)$ is the spectral representation of the input frame $x(t)$, R is the number of Harmonics to be considered and ω_i is the range of possible fundamental frequencies. $Y(\omega)$ is used to find the maximum value $\hat{Y}$ that is the fundamental frequency, as is shown in equation (2).

This method does not require huge computations; therefore, real-time processing is possible, but the low frequency resolution is limited by the number of evaluation samples. To overcome this disadvantage, longer evaluation samples need to be taken or the zero padding technique can be applied.

5.2. Autocorrelation

Autocorrelation function [3] shows the similarity of a signal to a lagged version of itself so that a pure periodic signal expresses periodic peaks in the function. Equation (3) shows the autocorrelation function.

$$R(n) = \frac{1}{N} \sum_{i=0}^{N-1} x(i)x(i+n) \tag{3}$$

If autocorrelation function measures the pitch of white noise, we can notice only one peak at zero lag (n=0) with small values for all other lags. Also, a periodic signal, such as a pure 100 Hz sine tone is measured, we can observe a peak at n=441 with sampling rate 44.1 kHz/s.

Autocorrelation is computationally expensive. To detect a frequency located between 20 Hz to 20 kHz, equation (3) shows that autocorrelation needs at least 2205 samples (n=2205) at sampling rate 44.1kHz/s. We take a window of the signal whose length is at least 2 times the period that we detect; therefore, finally the autocorrelation brings over 9 million computations (2205 × 4410) for each pitch measurement. From equation (3), we know that the number of computations for autocorrelation function is controlled by the lowest frequency that we detect.

The pitch of a human vocal signal occurs between 80 Hz and 900 Hz. We limit the lowest frequency that we detect from 20 Hz to 80 Hz to reduce the number of evaluation samples from 2205 to 600 and this computation reduction brings less than one million computations (600×1200). This reduction allows real-time processing of the proposed effecter on a 2 GHz Windows PC.

5.3. Comparison

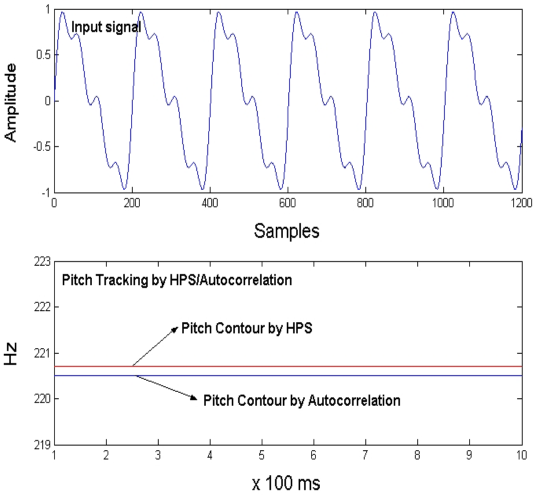


Figure 3 Complex Sine wave Pitch Contour

Figure 3 depicts the pitch detection results by HPS and autocorrelation function with a periodic input signal (4) with a pitch of 220 Hz at sampling rate 44.1 kHz/s.

$$y(t) = 0.8\sin(440\pi t) + 0.3\sin(880\pi t) + 0.1\sin(1320\pi t) + 0.2\sin(1760\pi t) \quad (4)$$

Figure 3 shows, both HPS and autocorrelation exhibit good performance within the frequency resolution of their own, but when the real vocal signal is applied, the results of two methods are totally different (Figure 4).

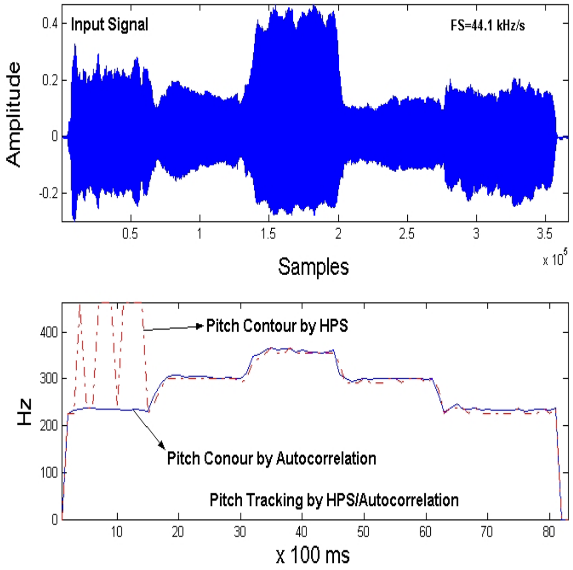


Figure 4 Vocal Signal Pitch Contour

The vocal input that is used for Figure 4 were five vowels (aha, eee, u, eh, oh) at different pitch. Figure 4 shows that the pitch detection using autocorrelation is relatively robust in a noisy environment compared to the other method.

The experimental result (Figure 4) demonstrates the accuracy of detection and robustness in noisy environments led us to use autocorrelation for our pitch detection method.

6. RESULT

The vocal signals are coming into the MIC/Line-In port and the vibrato effect is generated depending on the condition of the vocal signal. A pitch measurement is performed every 100 ms. Figure 5 depicts the output results generated by our vocal vibrato effecter. The same vocal input tested in 5.3 is applied and the three parameters mentioned earlier are: Latency Time 500 ms, Vibrato Depth 3 ms and Vibrato Speed 7 Hz.

From Figure 5, the vibrato is triggered after holding initial pitch for 500 ms of Latency Time along with defined Vibrato Depth and Vibrato Speed parameters without any external data support. According to our other experiments using the 'Vocal Vibrato Effecter', the results sounded natural with 7 Hz Vibrato Speed with 5 ms Vibrato Depth at Latency Time 300 ms for fast tempo songs and sounded natural with 5 Hz Vibrato Speed with 5 ms Vibrato Depth at Latency Time 700 ms for slow tempo songs.

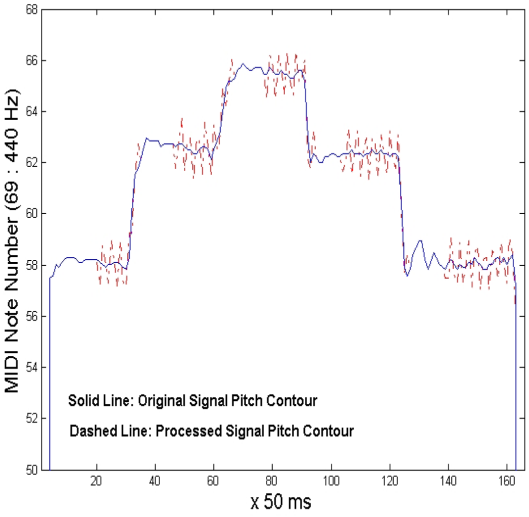
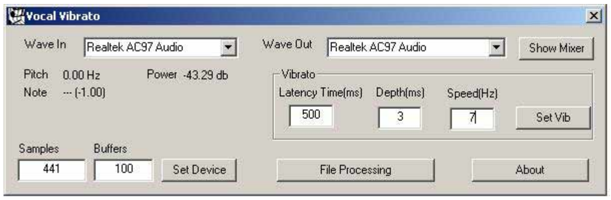


Figure 5 Pitch Tracking of Original Signal and Processed Signal



Wave In Realtek AC97 Audio

Wave Out Realtek AC97 Audio

Pitch 0.00 Hz

Power -43.29 db

Note --- (-1.00)

Vibrato

Latency Time(ms) 500

Depth(ms) 3

Speed(Hz) 7

Samples 441

Buffers 100

Figure 6 Vocal Vibrato Effecter Control Panel

7. CONCLUSION

We are able to demonstrate the ability to generate vocal vibrato effect intelligently with our proposed system. Also, this paper reveals the possibility to realize 'Vocal Vibrato Effecter' for Karaoke applications.

8. REFERENCES

[1] Ciara Leydon, Jay J. Bauer, and Charles R. Larson, "The role of auditory feedback in sustaining vocal vibrato", J.Acoust.Soc.Am, vol.114, no 3, pp. 1571-1581 (2003)

[2] Udo Zolzer, "DAFX Digital Audio Effects" Chapter 3 (2002)

[3] Perry R. Cook, "Real Sound Synthesis for Interactive Applications" Chapter 5 (2002)

[4] Patriciode la Cuadra, Patricio, Aaron Master and Craig Sapp, "Efficient Pitch Detection Techniques for Interactive Music" in the Proceedings of the International Computer Music Conference (ICMC) 2001, Havana, Cuba. pp. 403-406.

[5] Lau, Kim Hang, A System for Hybridizing Vocal Performance Preprint 5625; AES Convention 112; Preprint 5625, April 2002

[6] JON DATTORRO, "Effect Design Part2: Delay-Line Modulation and Chorus", J. Audio Eng Soc., Vol 45, No. 10, October 1997

[7] T. Mitsui and S. Hosokawa, "Karaoke Around the World" Introduction (1998)